RIAO'97 Montréal, Canada, 1997

# Le projet NADIA-DEC : vers un Dictionnaire Explicatif et Combinatoire informatisé ?

Gilles Sérasset

GETA-CLIPS-IMAG (UJF & CNRS)
BP 53
38041 Grenoble cedex 9
Tél. : 04.76.51.43.80 - Fax : 04.76.51.44.05
Courriel : Gilles.Serasset@imag.fr

## Introduction

Dans le domaine de l'ingénierie linguistique et de la connaissance, le problème des ressources lexicales et linguistiques s'est toujours posé. Néanmoins, l'avancée des techniques du Traitement Automatique des Langues Naturelles (TALN) l'a rendu plus sensible. Il nous faut maintenant pouvoir répondre à des besoins importants en terme de quantité, de qualité et de complexité. La complexité et la diversité des informations requises augmente avec les exigences des outils de TALN ainsi qu'avec le développement de nouvelles applications (humaines ou machinales). Si la récupération (semi)automatique d'information lexicale est une piste, elle ne pourra remplacer la création manuelle de dictionnaires.

Nous nous sommes donc intéressé à la construction d'outils pour lexicographes et lexicologues. Afin d'avoir une bonne compréhension des problèmes qui se posent, nous avons décidé d'informatiser un dictionnaire complexe, contenant de nombreuses informations structurées, le dictionnaire explicatif et combinatoire du français contemporain (DEC). Le DEC étant un travail de lexicologie, il ne s'agit donc pas à proprement parler d'un dictionnaire, mais plutôt d'un ensemble d'entrées destinées à illustrer une théorie linguistique. Ce ne sont donc pas les données que l'on a informatisé, mais le processus de rédaction de ces données.

Les travaux menés au cours du projet NADIA-DEC s'appuient d'une part sur le système SUBLIM [Sérasset 1994] défini au laboratoire GETA-CLIPS de l'université Joseph Fourier (Grenoble I) et, d'autre part, sur les travaux de lexicologie menés par l'équipe d'Igor Mel'čuk [Mel'čuk et al. 1995] au laboratoire GRESLET de l'université de Montréal. Il a reçu le soutient du réseau LTT de l'AUPELF-UREF et des ministères français et canadiens des affaires étrangères.

Ce projet pose des problèmes informatiques et linguistiques sérieux. L'aspect évolutif de la structure du DEC impose la nécessité de fournir des outils informatiques adaptables. La nécessité de formalisation des informations peut conduire à différentes stratégies de représentation des informations.

Nous montrons les différentes étapes de l'informatisation du DEC en donnant tout d'abord la structure interne des informations lexicales. Nous donnons ensuite un aperçu des outils et méthodes utilisées pour la création et la validation d'un DEC informatisé.

## Le projet NADIA-DEC

*Objectifs*

Depuis 1994, le GRESLET (université de Montréal) et le GETA-CLIPS (université Joseph Fourier - Grenoble I) travaillent ensemble sur le projet NADIA-DEC, soutenu par le réseau LTT de l'AUPELF-UREF.

L'objectif de ce projet est l'informatisation du Dictionnaire Explicatif et Combinatoire du français contemporain (DEC) créé par Igor Mel'čuk. Cette informatisation se base sur les travaux préalablement effectués au GETA ([Sérasset 1994]), et répond aux contraintes suivantes :

- **Fidélité linguistique** : Les structures informatiques utilisées doivent rester proches des structures linguistiques que l'on souhaite représenter,
- **Généricité** : Les outils construits doivent pouvoir être utilisés pour d'autres structures,
- **Adaptabilité** : Les outils informatiques doivent pourvoir évoluer en même temps que la structure informatique,

Le DEC étant en constante évolution, il nous est très vite apparu important d'informatiser non seulement les données existantes, mais surtout sa production. Nous avons donc développé des outils d'éditions spécialisés pour le DEC ainsi que des outils de récupération des informations déjà décrites disponibles sous forme de fichiers Word™.

L'approche utilisée ne remet pas en cause la structure linguistique que l'on peut trouver dans le DEC. La structure informatique du DEC doit permettre, au minimum, de re-générer à l'identique les fichiers Word™ utilisés pour la version papier. Aussi, toutes les informations sont présentes, et ce même si elles ne sont pas structurées. Il est ainsi toujours possible, au fur et à mesure que l'on avance dans ce projet, d'augmenter la structuration des données sans avoir a reprendre l'ensemble du processus de récupération à partir des fichier Word™.

À cet égard, le projet NADIA-DEC se distingue des autres projets d'informatisation du DEC, qui se basent a priori sur une structure informatique simplifiée et qui n'informatise que le sous ensemble de données commun entre le DEC et cette structure.

Enfin, les données du DEC ne sont pas récupérées dans le but d'une utilisation informatique particulière. Nous estimons que cette indépendance par rapport à l'usage qui sera fait des données nous permet de garantir la complétude des informations récupérées.

*Méthodologie*

Nous avons distingué plusieurs tâches pour accomplir le projet NADIA-DEC :

- définition d'une structure informatique pour le DEC,
- récupération des informations existantes sous cette forme structurée,
- construction d'un éditeur spécialisé pour cette structure (l'éditeur DECID),
- exporter les données structurées vers différentes formes.

Ainsi, notre méthodologie peut être résumée par le schéma suivant :

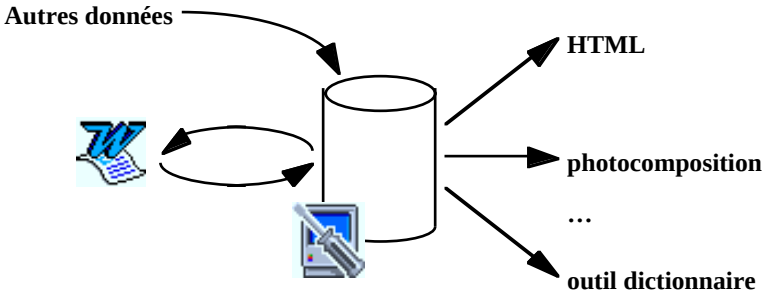


Figure 1 : Méthodologie de création d'un DEC informatisé.

## L'éditeur DECID

DECID est un éditeur spécialisé pour l'édition du DEC. Il offre de nombreuses fonctionnalités pour aider le lexicographe. Sa conception et son implantation ont été effectuées dans un soucis de simplicité et de convivialité.

En utilisant l'éditeur DECID, le lexicographe créé ou modifie, en direct, une structure informatique. Pourtant, l'interface a été conçue pour lui donner l'impression de travailler, comme auparavant, sur le DEC tel qu'il est publié. Aussi, la visualisation des données est-elle très proche de celle qui est utilisée dans la version papier.

Le lexicographe dispose d'une fenêtre principale lui donnant la liste des vocables et des lexies du fichier en cours d'édition (figure 2). Le second type de fenêtre présente et permet d'éditer une lexie.

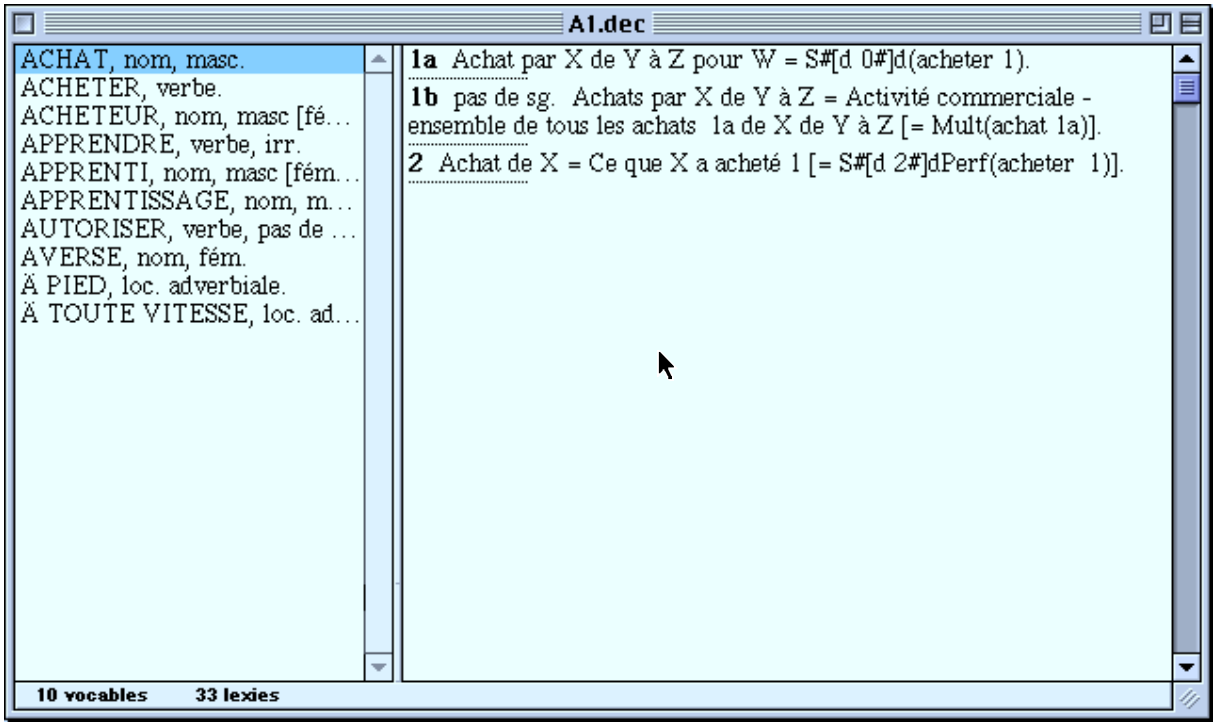


Figure 2 : La fenêtre principale. La zone de gauche présente la liste des vocables du fichier en cours d'édition. La zone de droite présente la liste des lexie du ou des vocables sélectionnés.

La fenêtre de lexie permet d'éditer le vocable, le numéro de la lexie, les informations morphologiques, la définition et les exemples de manière très simple. La zone de régime n'est pas encore traitée par l'éditeur DECID.

Les fonctions lexicales apparaissent sous une forme très proche de la forme papier. Mais leur édition a été rendue très aisée par l'éditeur DECID. En effet, auparavant, le lexicographe devait, pour éditer une fonction lexicale, utiliser des indices, des exposants, des changements de fontes. Il devait faire attention a la correction du nom de la fonction, bien mettre le premier caractère en majuscule et le reste en minuscule… Tous ces soucis ont maintenant disparu lorsqu'on utilise l'éditeur DECID. En effet, le lexicographe se contentera de taper : `perm1incepreal3+'usual` pour voir se dessiner la fonction: $\mathbf{Perm}_1\mathbf{IncepReal}_{3\subset}^{\text{usual}}$.

Le lexicographe pourra ensuite sauver son travail sous forme structurée ou alors sous forme d'un fichier RTF (Rich Text Format) qu'il pourra ensuite utiliser directement en Word.

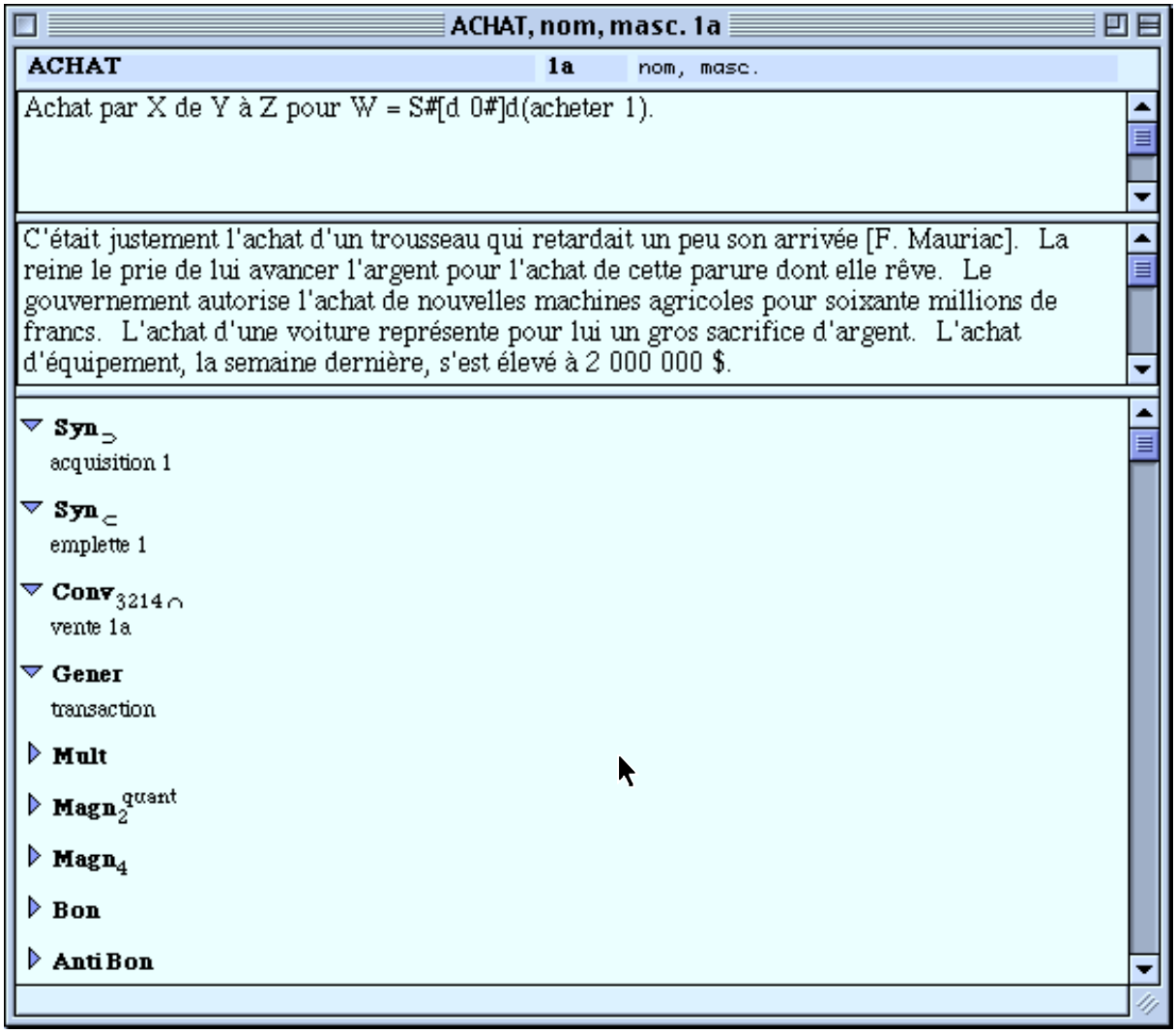


Figure 3 : La fenêtre de lexie pour la lexie achat 1a. En plus des zones de définition et d'exemple, on remarque la zone des fonctions lexicales.

## Récupération des données existantes

En plus de l'éditeur DECID, nous avons développé un outil de récupération des données publiées du DEC. Cet outil part d'un fichier en RTF (Rich Text Format) généré à partir des fichier Word™ qui ont été utilisés à l'origine pour la création du DEC papier.

Cette récupération n'a pu se faire que semi-automatiquement. Les fichiers en cours de récupération devant être corrigés pour être récupérable. Fort heureusement, les fichiers avaient, dès le départ été créés en utilisant des styles cohérents pour les différents paragraphes décrivant une entrée (définition, régime, etc.). Sans cela, la récupération n'aurait pu avoir lieu.

Certaines difficultés sont dues à l'outil Word™ utilisé. L'absence totale de documentation du format RTF nous a obligé à produire des outils ad-hoc. De plus, pour des raisons encore assez obscures, des fichiers d'apparence identiques ont des descriptions RTF différentes. Ainsi, deux paragraphes successifs ayant le même style peuvent apparaître soit comme deux paragraphes indépendants (l'information de style est donnée pour chaque paragraphe) ou comme deux paragraphes identique (la définition de style n'est donnée qu'au début du premier).

D'autres difficultés sont dues au DEC lui même. Le DEC a été conçu au départ sous une forme papier utilisable par un homme. Aussi, le DEC a été conçu avant tout par sa **présentation**. Aussi, une très grande importance a été accordée à la forme plus qu'à la structure. Ainsi, certaines erreurs dans les documents Word n'étaient pas détectées car elle n'étaient pas visible sur papier. Par exemple, la définition et la liste d'exemples ont une même

forme, mais sont représentées par deux styles différents. Néanmoins, on trouve souvent des erreurs dues à l'identité de forme de ces deux types d'information, l'humain pouvant très facilement faire la différence par le contexte.

Dans les tableaux de régimes la présentation était contrôlée entièrement par le lexicographe. Ainsi, certaines lignes pouvaient être séparées par une marque de fin de paragraphe, un saut à la ligne, ou même par une succession de tabulations. Certaines valeurs pouvaient tenir sur plusieurs lignes. Dans ce cas, les valeurs apparaissent, dans le fichier séquentiel RTF, de manière entrelacée.

Ces erreurs de présentation et d'édition ont été réglées. Elles ont été fort heureusement assez mineures.

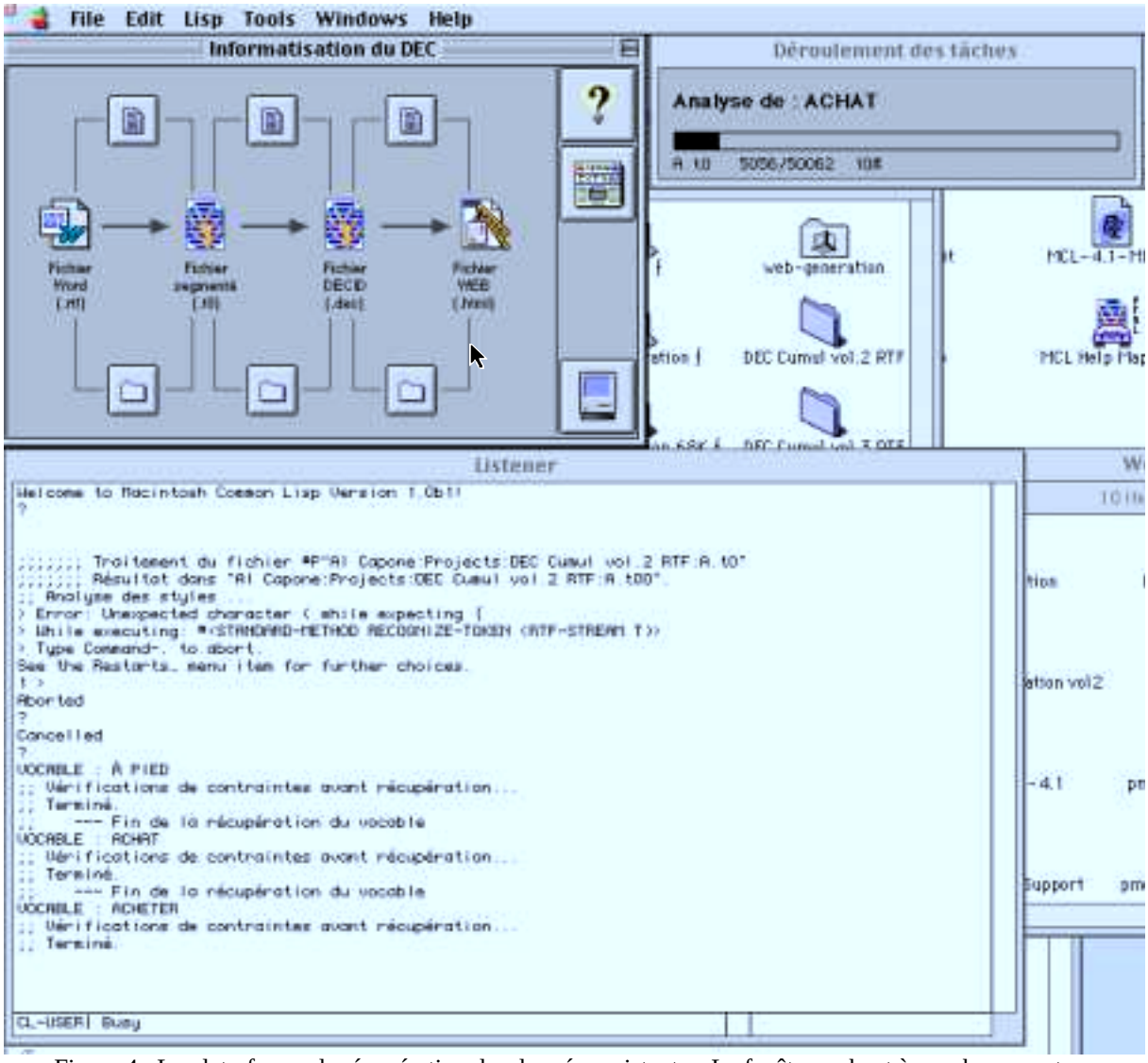


Figure 4 : La plate-forme de récupération des données existantes. La fenêtre en haut à gauche permet au récupérateur de déclencher les traitements. La fenêtre du bas donne des informations sur le traitement en cours.

En utilisant la plate-forme de récupération (figure 4), le lexicographe déclenche la récupération d'un fichier RTF ou de tous les fichiers RTF présents dans un dossier. Ces fichiers sont analysés puis, pour chaque vocable détecté, un fichier est créé qui contient les informations sous forme structurés. Pour chaque vocable, le lexicographe peut demander un diagnostic qui lui sera utile pour savoir si les données ont été récupérées de manières satisfaisantes ou si une erreur s'est glissée dans la récupération sans que le processus ne se soit interrompu (figure 5).

L'outil de récupération aide le correcteur en donnant un diagnostic du vocable récupéré. Ainsi, par comparaison avec la version papier, il est aisé de voir ce qui n'a pas fonctionné et pourquoi.

```
;; Vocable      : ACHAT
;; Catégorie    : nom, masc.
;; Possède un tableau résumé : 3 résumés.
;; Pas de note.
;;   Lexie      : 1a.
;;    Pas de connotations.
;;    Possède des informations de régime :
;;     Tableau à 4 colonne(s)
;;     1 = X | 2 = Y | 3 = Z | 4 = W
;;       3   |   1   |   2   |   2
;;     3 restriction(s) numérotées.
;;     6 exemples de réalisations
;;    Possède des fonctions lexicales :
;;      • 11 fonction(s) lexicale(s).
;;    Possède des exemples.
;;   Lexie      : 1b.
;;    Pas de connotations.
;;    Possède des informations de régime :
;;     Tableau à 3 colonne(s)
;;     1 = X     |    2 = Y  |    3 = Z
;;       3       |      2    |      1
;;     2 restriction(s) numérotées.
;;     2 exemples de réalisations
;;    Possède des fonctions lexicales :
;;      • 14 fonction(s) lexicale(s).
;;    Possède des exemples.
;;   Lexie      : 2.
;;    Pas de connotations.
;;    Possède des informations de régime :
;;     Tableau à 1 colonne(s)
;;     2 = X
;;       2
;;     0 restriction(s) numérotées.
;;     1 exemples de réalisations
;;    Possède des fonctions lexicales :
;;      • 3 fonction(s) lexicale(s).
;;    Possède des exemples.
```

Figure 5 : Le diagnostic de récupération d'un vocable.

## Exploitation des données informatisées

Les données ainsi informatisées ont été exportées sous forme HTML. Nous avons ainsi pu produire automatiquement un site web complet présentant le DEC sous une **présentation** analogue à celle utilisées dans la version papier. Les figures 6 et 7 représentent une page du DEC vue par un navigueur standard.

Cette version HTML du DEC est, comme le DEC au format Word, destinée à un usage humain. Nous avons adopté une forme aussi proche que possible de la forme originale. Mais cette présentation pose différents problèmes. En effet, le format HTML ne permet pas, de manière simple de préciser effectivement une forme. C'est le navigueur, qui, en dernier ressort effectue la présentation. Cela pose différents problèmes :

- les tables, n'apparaissent pas de la même manière suivant les navigueur utilisés. Cela peut mener à des colonnes trop larges ou trop étroites,
- tous les navigueurs ne savent pas forcément interpréter et présenter les informations en indice ou en exposant,

Enfin, indépendamment de la compatibilité des différents navigueurs, la version HTML du DEC pose des problème intrinsèques. Ainsi, certains caractères sont propres au DEC (ex : k et l qui délimitent les locutions ou ' et " qui délimitent les sémantèmes). Ces caractères ne sont présents dans aucune fonte standard. Actuellement, HTML ne permet pas d'inclure une fonte ou une description de caractère qui soit portable. On peut indiquer au navigueur d'utiliser une fonte particulière, mais celle-ci doit être présente dans le système du client. L'utilisation d'une image pour ces caractères peut être envisagée, mais le client n'aura pas une bonne présentation s'il décide de changer la taille des caractères affichés (l'image ne grandira pas en fonction).

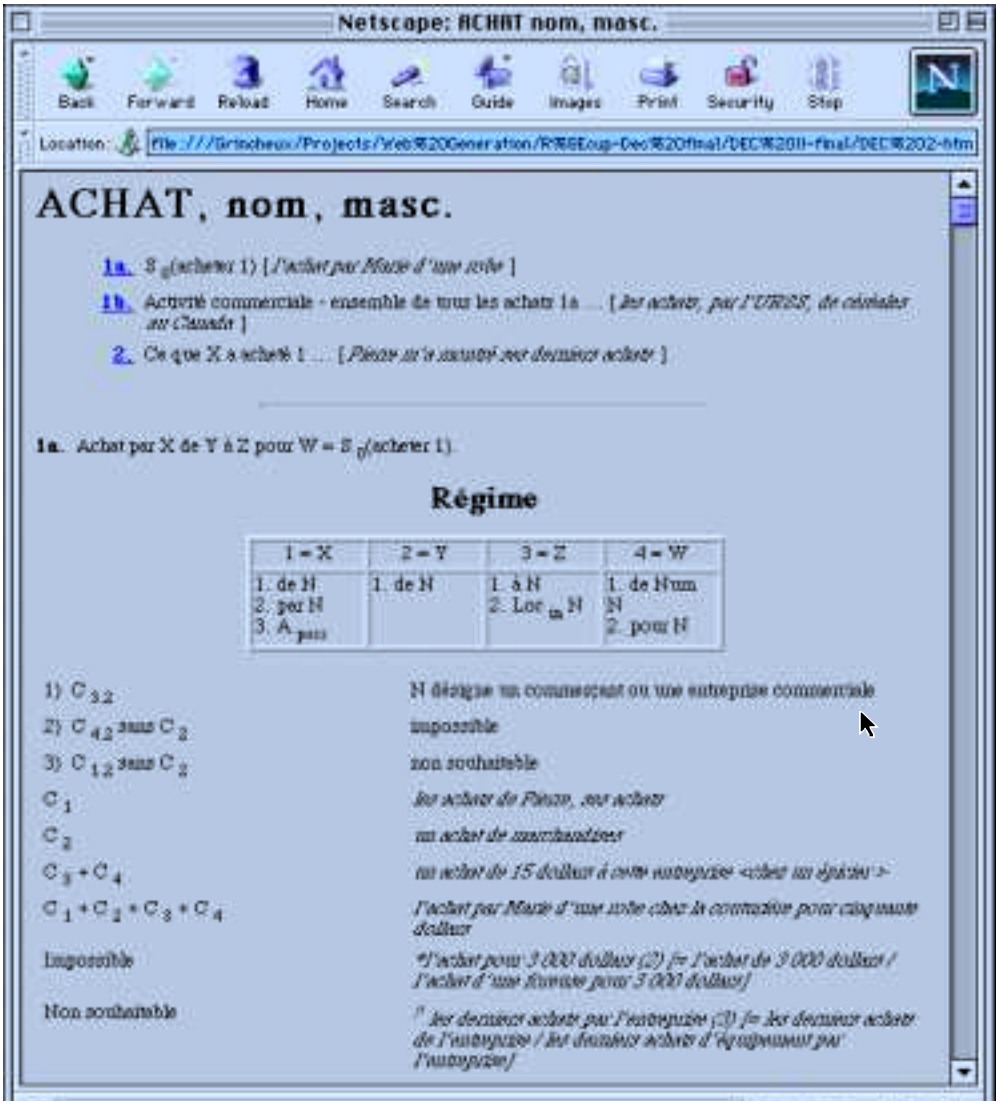


Figure 6 : Le vocable ACHAT nom, masc vu par un navigateur standard.

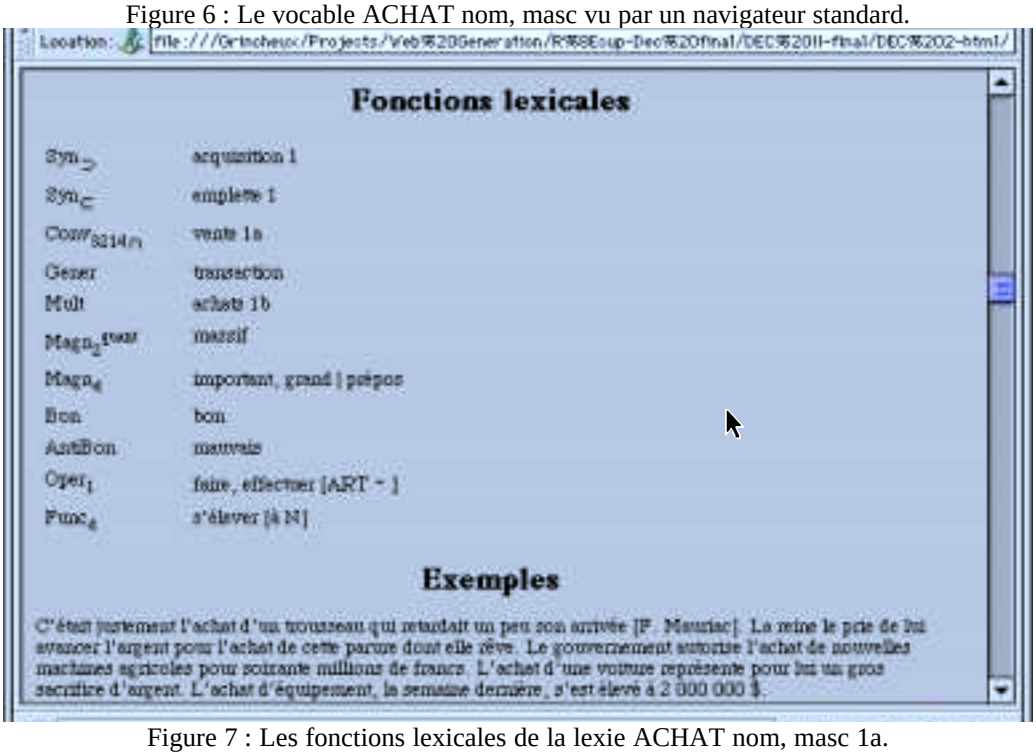


Figure 7 : Les fonctions lexicales de la lexie ACHAT nom, masc 1a.

## Conclusion

À l'occasion de l'action de recherche partagée NADIA-DEC, nous avons donc pu informatiser le dictionnaire explicatif et combinatoire du français contemporain. Pour cela, nous avons récupéré semi automatiquement la totalité des entrées des volumes II et III du DEC.

Nous avons aussi créé un outil d'édition spécialisé pour le DEC. Cet outil apporte des avantages certains au lexicographes, mais il contraint trop la structure du dictionnaire et ne pourra être utilisé tant que l'édition du DEC se fera dans l'optique d'une recherche en lexicologie (structure en cours de définition).

Néanmoins, les travaux effectués seront très utiles pour un passage en phase de production du DEC ou d'un dictionnaire dérivé.

Ce travail a constitué une étape importante dans nos recherches sur des outils pour lexicographes. Elle nous a permis de mettre en œuvre nos méthodes sur un dictionnaire très complexe. Nous avons ainsi pu valider certains choix. Néanmoins, nous avons pu voir que la construction d'outils spécialisés fige la structure informatique utilisée. Or dès que l'on travail avec des dictionnaires assez compliqué, la possibilité de remise en cause des structures informatiques en cours d'édition d'un dictionnaire est nécessaire.

Aussi, nous souhaitons orienter nos recherches sur des méthodes génériques de création d'outils spécialisés pour lexicographes. Cette généricité nous permettra d'offrir des outils évolutifs et rendra plus facile les recherches de lexicologie.

## Bibliographie

Igor Mel'čuk, André Clas et Alain Polguère (1995). *Introduction a la lexicologie explicative et combinatoire* Universités francophones et champs linguistiques, Louvain-la Neuve, AUPELF-UREF et Duculot.

Gilles Sérasset (1994). *SUBLIM : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions*, Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1 : 194 p.

Gilles Sérasset (1995). *Informatisation du Dictionnaire Explicatif et Combinatoire : le projet NADIA-DEC*, Lexicomatique et Dictionnairique, Lyon, 28-30 septembre 1995, pp. 205-215.

Gilles Sérasset (1996). *Un éditeur pour le dictionnaire explicatif et combinatoire du français contemporain*, Journées lexique du PRC-CHM, Grenoble, 13-14 novembre 1996, pp. 131-138.

Gilles Sérasset (1997). *Informatisation du dictionnaire Explicatif et Combinatoire*, TALN'97, Grenoble, 12-13 juin 1997, pp. 194-198

Gilles Sérasset et Étienne Blanc (1993). *Une approche par acceptions pour les bases lexicales multilingues*, T-TA-TAO 93, Montréal, 30 septembre-2 octobre 1993, pp. 65-84.

Gilles Sérasset et Alain Polguère (1997). *Outils pour lexicographes : application à la lexicographie explicative et combinatoire*, RIAO'97, Montréal, 25-27 juin 1997, pp. 701-708.